THE UNITED STATES PATENT AND TRADEMARK OFFICE

Application No. : 09/835,064)
) Primary Examiner: Abel Jalil, Neveen
Filing Date : 04/13/2001)
) Group Art unit: 2165
First Named Inventor : G. GIUFFRIDA)
)
Firm Docket No. : HRL065)
)
For: A Method And Apparatus For)
Automatically Extracting Metadata From)
Electronic Documents Using Spatial Rules)
______________________________)

37 C.F.R. 1.131 DECLARATION

I, Dan Allemeier, do hereby declare that:

1. The inventors of the invention described in non-provisional utility application serial no. 09/835,064, filed on 04/13/2001, and entitled "A Method And Apparatus For Automatically Extracting Metadata From Electronic Documents Using Spatial Rules" are unavailable to sign the attached declaration made under 37 CFR 1.131.

2. The assignee or other party in interest may make such a declaration when it is not possible to produce the affidavit or declaration of the inventor. *Ex parte Foster*, 1903 C.D. 213, 105 O.G. 261. As such the present declaration under 37 CFR 1.131 has been made under the provisions of 715.04, subsection D of the MPEP, which allows the Assignee to make the Declaration when the inventors are unavailable.

3. The non-provisional utility application serial no. 09/835,064, filed on 04/13/2001, and entitled "A Method And Apparatus For Automatically Extracting Metadata From Electronic Documents Using Spatial Rules" was assigned to HRL

Laboratories, LLC, having a place of business located at 3011 Malibu Canyon Road, Malibu, CA 90265. The assignment was officially recorded by the USPTO on Reel/Frame 013125/0430, having a date of recordation of 7/26/2002.

4. I am the General Counsel of HRL Laboratories, LLC. As the General Counsel, I have the authority to make the following statements on behalf of the Assignee.

5. I have knowledge of the contents of the invention described in non-provisional utility application serial no. 09/835,064, filed on 04/13/2001, and entitled "A Method And Apparatus For Automatically Extracting Metadata From Electronic Documents Using Spatial Rules." The invention as described in the aforementioned non-provisional utility application was invented at least as early as January of 1999, and by acts undertaken wholly in the United States of America, the invention was diligently pursued with the purpose of its actual reduction to practice at least as early as May of 1999.

6. Appendix A is an internal document restricted from dissemination outside of the company. It is a copy of the original invention disclosure statement submitted to the legal counsel of HRL Laboratories, LLC and was received on February 20, 2000. The document includes a total of sixteen (16) pages and presents the concepts disclosed and claimed in the present application. Each sheet was signed by the inventors on or before February 10, 2000. The signatures of each of the inventors are located at the bottom of each sheet. Each of the sixteen (16) pages is date stamped in the bottom left hand corner of the sheet. The document establishes proof of concept and a reduction to practice of the present invention, in its entirety, and as arranged in the claims of the present application, at least as early as May of 1999. This is corroborated by the "Reduction To Practice" table on sheet 2 of Appendix A in Section 4.

7. As stated above, the contents of the present application, as arranged in the claims, were reduced to practice at least as early as May of 1999. The specific details on an element-by-element basis for each of the independent claims, Claims 1 and 9, are as follows:

    a. Claim 1 as previously presented claims an apparatus for automatically extracting metadata from electronic documents comprising a first processing element, a second processing element, a reasoning element, and a database, wherein: said first processing element is configured to convert electronic documents into files; said first processing element is configured to provide the files to a second processing element; said second processing element is configured to receive said files and extract predetermined information from the files; said second processing element is further configured to provide said extracted predetermined information to said reasoning element; said database is configured to provide input to said reasoning element; said reasoning element is configured to employ a set of rules to automatically extract metadata from the files by employing the extracted predetermined information and the input from the database; and said reasoning element provides an output of metadata.

        i. The element "a first processing element, a second processing element, a reasoning element, and a database wherein: said first processing element is configured to convert electronic documents into files" was conceived at least as early as May of 1999. The conversion process of the electronic documents into files is represented by the diagram of Figure 1 and is described in Section 9, sheets 7-14.

        ii. The element "said first processing element is configured to provide the files to a second processing element; said second processing element is configured to receive said files and extract predetermined information from the files" was conceived at least as early as May of 1999. The movement of the files from the first processing element to a second processing element is represented by the diagram of Figure 1 and is described in Section 9, sheets 7-14.

iii. The element "said second processing element is further configured to provide said extracted predetermined information to said reasoning element" was conceived at least as early as May of 1999. The movement of the files from the first processing element to a second processing element is represented by the diagram of Figure 1 and is described in Section 9, sheets 7-14.

iv. The element "said reasoning element is configured to employ a set of rules to automatically extract metadata from the files by employing the extracted predetermined information and the input from the database" was conceived at least as early as May of 1999. The movement of the files from the first processing element to a second processing element is represented by the diagram of Figure 1 and is described in Section 9, sheets 7-14.

v. The element "said reasoning element provides an output of metadata" was conceived at least as early as May of 1999. The reasoning element providing an output of metadata is described in Section 9, sheets 7-14. The concept was reduced to practice at least as early as May of 1999, the experimental results of which are discussed on sheets 9-14.

b. Claim 9 claims a method for automatically extracting metadata from electronic documents providing a first processing element, a second processing element, a reasoning element, and a database and comprising the steps of: employing said first processing element to convert electronic documents to files; further employing said first processing element to provide the files to said second processing element; employing said second processing element to receive said files and extract predetermined information from the files; further employing said second processing element to provide extracted predetermined information to said reasoning element; employing said database to provide input to said reasoning element; employing a set of rules in said reasoning element to automatically extract metadata from the files by employing the extracted predetermined information and the input from the database; and providing an out put of metadata from said reasoning element.

i. The act of "employing said first processing element to convert electronic documents to files" was conceived at least as early as May of 1999. The conversion process of the electronic documents into files is represented by the diagram of Figure 1 and is described in Section 9, sheets 7-14.

ii. The act of "further employing said first processing element to provide the files to said second processing element" was conceived at least as early as May of 1999. The movement of the files from the first processing element to a second processing element is represented by the diagram of Figure 1 and is described in Section 9, sheets 7-14.

iii. The act of "employing said second processing element to receive said files and extract predetermined information from the files" was conceived at least as early as May of 1999. The movement of the files from the first processing element to a second processing element is represented by the diagram of Figure 1 and is described in Section 9, sheets 7-14.

iv. The element "further employing said second processing element to provide extracted predetermined information to said reasoning element" was conceived at least as early as May of 1999. The movement of the files from the first processing element to a second processing element is represented by the diagram of Figure 1 and is described in Section 9, sheets 7-14.

v. The element "employing said database to provide input to said reasoning element" was conceived at least as early as May of 1999. The reasoning element providing an output of metadata is described in Section 9, sheets 7-14. The concept was reduced to practice at least as early as May of 1999, the experimental results of which are discussed on sheets 9-14.

vi. The element "employing a set of rules in said reasoning element to automatically extract metadata from the files by employing the extracted predetermined information and the input from the database" was conceived at least as early as May of 1999. The reasoning element providing an output of metadata is described in Section 9, sheets 7-14. The concept was reduced to practice at least as early as May of 1999, the experimental results of which are discussed on sheets 9-14.

vii. The element "providing an out put of metadata from said reasoning element" was conceived at least as early as May of 1999. The reasoning element providing an output of metadata is described in Section 9, sheets 7-14. The concept was reduced to practice at least as early as May of 1999, the experimental results of which are discussed on sheets 9-14.

8. I hereby declare that all statements made herein of our own knowledge are true and that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under 18 U.S.C. 1001 and that such willful false statements may jeopardize the validity of the application or any patent issued thereon.

3|5|2008

Dan Allemeier Date

SEND COMPLETED DISCLOSURE DIRECT TO:
HRL LABORATORIES PATENTS AND
LICENSING

INVENTION DISCLOSURE

THIS INVENTION DISCLOSURE IS [illegible]E
PURSUANT TO MY / OUR INVENTION AG[illegible]EMENT
WITH HRL LABORATORIES, LLC.

HRL065

LABORATORIES

1

SHEET 1 OF ______

1. TITLE OF INVENTION

Method for Automatically Extracting Metadata from Electronic Documents Using Spatial Rules

2. INVENTOR(S)

NAME	PAYROLL NO.	SOURCE CODE	LOC	BLDG	MS	PHONE	MANAGER
Giovanni Giuffrida	Vendor	30-21-40	MA	254	RL96	X5317	San Dao
Eddie Shek	F6554	30-21-40	MA	254	RL96	X5607	San Dao
Jihoon Yang	J3313	30-21-40	MA	254	RL96	X5505	San Dao

This is to acknowledge that the above Invention Disclosure has been received by HRL Laboratories, LLC Patents and Licensing. The disclosure will be reviewed at the next Evaluation Committee Meeting of your organization and you will be promptly informed of the results. If you have any questions please contact the patent attorney listed on the bottom of this form.

This sheet will be returned to the inventor(s) as a confirmation of receipt by HRL Laboratories, LLC Patents and Licensing.

LOSS OF RIGHTS THROUGH RELEASE TO THE PUBLIC

The right to apply for and obtain a valid patent may be lost as the result of certain activities, such as (1) disclosing the invention outside of the company without an appropriate confidentiality agreement with the receiving party; (2) using the invention publicly; (3) using the invention privately to build or test items that are to be sold publicly; or (4) putting the invention "on sale" by selling or offering for sale an item or product that embodies or uses the invention, or is made or tested by use of the invention. Submitting a proposal with the intent to use the invention in the performance of a resulting contract puts the invention "on sale." Please inform me immediately of any of these activities or any plans to undertake any of them.

ASSIGNED ATTORNEY: ______________________ PHONE ( ) ______________________

HRL LABORATORIES PROPRIETARY

THIS DOCUMENT CONTAINS PROPRIETARY INFORMATION, AND EXCEPT WITH WRITTEN PERMISSION OF HRL LABORATORIES, LLC, SUCH INFORMATION SHALL NOT BE PUBLISHED, OR DISCLOSED TO OTHERS, OR USED FOR ANY PURPOSE, AND THE DOCUMENT SHALL NOT BE DUPLICATED IN WHOLE OR IN PART. THIS LEGEND SHALL BE APPLIED TO ALL DOCUMENTS CONTAINING THIS INFORMATION.

PATENT DOCKET NO.

SIGNATURE INVENTOR — DATE 1/27/2000

SIGNATURE INVENTOR — DATE 2/10/2000

SIGNATURE INVENTOR — DATE 2/10/2000

SIGNATURE INVENTOR — DATE

APPENDIX A

SEND COMPLETED DISCLOSURE DIRECT TO:
HRL LABORATORIES PATENTS AND
LICENSING

INVENTION DISCLOSURF

THIS INVENTION DISCLOSURE IS I.   E
PURSUANT TO MY / OUR INVENTION AGRcEMENT
WITH HRL LABORATORIES, LLC.

HRL LABORATORIES

SHEET 2 OF ______

Method for Automatically Extracting Metadata from Electronic Documents Using Spatial Rules

2. INVENTOR(S)

NAME	PAYROLL NO.	SOURCE CODE			LOC	BLDG	MS	PHONE	MANAGER
Giovanni Giuffrida	Vendor	30	21	40	MA	254	RL96	X5317	San Dao
Eddie Shek	F6554	30	21	40	MA	254	RL96	X5607	San Dao
Jihoon Yang	J3313	30	21	40	MA	254	RL96	X5505	San Dao

A. BY WHOM WAS FIRST DESCRIPTION WRITTEN OR DRAWING MADE?	DATE	TIME SPENT	ACCT. CHARGED	LOCATION OF FIRST DESCRIPTION / DRAWING
Giovanni Giuffrida	1/99	4 Months	CD192A6QL	HRL
B. TO WHOM WAS INVENTION FIRST DISCLOSED?	DATE			
Jihoon Yang	2/99			

4. REDUCTION TO PRACTICE

A. WAS A DEVICE EMBODYING THE INVENTION CONSTRUCTED AND TESTED OR THE PROCESS PRACTICED?	YES X NO	BY WHOM Inventor	DATE STARTED 2/99	DATE COMPLETED 5/99	TIME SPENT 3 Months

B. ACCOUNT CHARGED — TIME	ACCOUNT CHARGED — MATERIAL	PRESENT LOCATION OF DEVICE
CD192A6QL		HRL

C. PRESENT LOCATION OF DOCUMENTS (DATE SIGNED AND WITNESSED), INCLUDING PHOTOS, DRAWINGS, AND DATA SHEETS SHOWING REDUCTION TO PRACTICE

NOTE: ALL EVIDENCE OF CONCEPTION (FIRST DRAWING AND FIRST WRITTEN DESCRIPTION) AND EVIDENCE OF REDUCTION TO PRACTICE (DEVICE EMBODYING THE INVENTION AND TEST DATA) MUST BE RETAINED.

5. RELATION TO GOVERNMENT CONTRACT

A. DOES THIS INVENTION RELATE TO WORK PERFORMED UNDER A GOVERNMENT CONTRACT?	YES NO X	CONTRACT NUMBER AND TITLE
B. IS INVENTION BEING USED ON A GOVERNMENT CONTRACT?	YES NO X	CONTRACT NUMBER AND TITLE

6. RELATED DOCUMENTS AND DISCLOSURE (BY YOU OR BY ANOTHER). PLEASE ATTACH COPY.

A. IS THERE A PUBLICATION OR PUBLIC PRESENTATION RELATED TO THE INVENTION?	YES NO X	DATE	IDENTIFY
B. ARE THERE ANY RELATED INVENTION DISCLOSURES OR PATENT APPLICATIONS?	YES NO X	DATE	IDENTIFY PD NO. ETC.
C. ARE THERE ANY PROPOSALS OR REPORTS OR OTHER DOCUMENTS RELATING TO THIS INVENTION	YES NO X	DATE	IDENTIFY
D. HAS THE INVENTION BEEN USED, DISCUSSED, DEMONSTRATED OR OTHERWISE DISCLOSED OUTSIDE THE COMPANY (SUCH AS TO A VENDOR OR CUSTOMER)?	YES NO X	DATE	TO / FOR WHOM (COMPANY / PERSON)

7. SALE

A. HAS PRODUCT EMBODYING INVENTION OR MADE BY INVENTION BEEN PROPOSED, SOLD, OR OFFERED FOR SALE?	YES NO X	ORDER NO.	ORDER DATE	DELIVERY DATE	DATE OFFERED OR PROPOSED

HRL LABORATORIES PROPRIETARY

THIS DOCUMENT CONTAINS PROPRIETARY INFORMATION, AND EXCEPT WITH WRITTEN PERMISSION OF HRL LABORATORIES, LLC, SUCH INFORMATION SHALL NOT BE PUBLISHED, OR DISCLOSED TO OTHERS, OR USED FOR ANY PURPOSE, AND THE DOCUMENT SHALL NOT BE DUPLICATED IN WHOLE OR IN PART. THIS LEGEND SHALL BE APPLIED TO ALL DOCUMENTS CONTAINING THIS INFORMATION

SIGNATURE INVENTOR	1/27/2000 DATE
SIGNATURE INVENTOR	2/10/2000 DATE
SIGNATURE INVENTOR	2/10/2000 DATE
SIGNATURE INVENTOR	DATE

READ AND UNDERSTOOD BY:

Darrel J Van Buer WITNESS NAME (TYPE)	SIGNATURE	2/21/2000 DATE
Wensheng Zhou WITNESS NAME (TYPE)	SIGNATURE	2/21/2000 DATE

PATENT DOCKET NO.

PD# 000211

(EACH PAGE UPON WHICH INFORMATION IS ENTERED SHOULD BE SIGNED AND WITNESSED)

SEND COMPLETED DISCLOSURE DIRECT TO:
HRL LABORATORIES PATENTS AND LICENSING

INVENTION DISCLOSURE

THIS INVENTION DISCLOSURE IS MADE
PURSUANT TO MY / OUR INVENTION AGREEMENT
WITH HRL LABORATORIES, LLC.

HRL LABORATORIES 2

SHEET 3 OF ______

B. IS PRODUCT EMBODYING INVENTION OR MADE BY INVENTION IN A DELIVERABLE ITEM?	YES NO X	DELIVERY DATE

HRL LABORATORIES PROPRIETARY

THIS DOCUMENT CONTAINS PROPRIETARY INFORMATION, AND EXCEPT WITH WRITTEN PERMISSION OF HRL LABORATORIES, LLC, SUCH INFORMATION SHALL NOT BE PUBLISHED, OR DISCLOSED TO OTHERS, OR USED FOR ANY PURPOSE, AND THE DOCUMENT SHALL NOT BE DUPLICATED IN WHOLE OR IN PART. THIS LEGEND SHALL BE APPLIED TO ALL DOCUMENTS CONTAINING THIS INFORMATION.

SIGNATURE INVENTOR — DATE 1/27/2000

SIGNATURE INVENTOR — DATE 2/10/2000

SIGNATURE INVENTOR — DATE 2/10/2000

SIGNATURE INVENTOR — DATE

READ AND UNDERSTOOD BY:

Darrel J VanBuer — WITNESS NAME (TYPE) — SIGNATURE — DATE 2/21/2000

Wensheng Zhou — WITNESS NAME (TYPE) — Wensheng Zhou SIGNATURE — DATE 2/21/2000

PATENT DOCKET NO.

(EACH PAGE UPON WHICH INFORMATION IS ENTERED SHOULD BE SIGNED AND WITNESSED)

SEND COMPLETED DISCLOSURE DIRECT TO:
HRL LABORATORIES PATENTS AND
LICENSING

INVENTION DISCLOSURE

THIS INVENTION DISCLOSURE IS MADE
PURSUANT TO MY / OUR INVENTION AGREEMENT
WITH HRL LABORATORIES, LLC.

HRL LABORATORIES

8. SUMMARY OF THE INVENTION

A. GIVE A BRIEF DESCRIPTION OF YOUR INVENTION, PARTICULARLY POINTING OUT WHAT IS BELIEVED TO BE NOVEL (THE "HEART" OF WHAT IS NEW).

Our invention uses a *spatial knowledge based* approach for the automatic extraction of metadata from electronic documents in preparation for archival in or dissemination through a digital library system. Our invention is based on mimicking the *visual/spatial knowledge* humans make use of when reading a document. In general, within a document category, a certain visual layout can be identified for all documents of that category. For instance, in a scientific paper it is usually the case that the title is located on the *upper* portion of the *first page* and it is printed using the *largest font* on the *first page*, and authors are listed *immediately under* the title in some order. Rules encoding the above and other common spatial layout properties are captured and embedded in a knowledge based system to model this type of reasoning to facilitate the extraction of metadata from the input document. A widely-used rule-based language, Postscript, that encodes these spatial facts for typesetting is used as the intermediate language for electronic documents in the system that currently realizes our invention concept.

B. EXPLAIN THE PURPOSE AND ADVANTAGES OF YOUR INVENTION. (WHAT WILL THE INVENTION DO BETTER THAN DONE PREVIOUSLY?)

The recent proliferation of computers and communication networks has made it possible for individuals around the world to access a wide variety of information sources through the Internet. Such information sources are fairly diverse and range from simple text (e.g., journal articles, conference papers, broadcast news) to complex multimedia (e.g., video, audio, image) data. The design of sophisticated tools for accessing these information sources and extracting knowledge from the data is thus of great interest and importance.

Digital libraries have been introduced in the Internet to store a variety of documents and to provide services with the documents. These documents include, among others, journal articles, conference papers, technical reports, and dissertations. Most of these digital libraries retrieve relevant documents by a keyword-based search in human-generated database indices. Our invention is of great help for the automatic generation of such indices.

C. IDENTIFY THE COMPANY PROGRAM OR PRODUCT LINE TO WHICH THE INVENTION APPLIES, AND THE EXPECTED VALUE TO THE PROGRAM OR PRODUCT LINE. ALSO IDENTIFY POTENTIAL COMMERCIAL APPLICATION OF THIS INVENTION, INCLUDING AUTOMOTIVE APPLICATIONS, IF ANY.

The invention represents a fundamental capability necessary in the construction on large-scale digital library systems that aim to support the gathering, searching, and dissemination of electronic documents such as those accessible through the WWW.

D. IDENTIFY THE PRIOR ART KNOWN TO YOU WHICH IS IMPROVED UPON OR DISPLACED BY YOUR INVENTION, AND STATE IN DETAIL, IF KNOWN, THE DISADVANTAGES OF THE CLOSEST PRIOR ART.

Prior systems that aim to automatically extract descriptive information from electronic documents do not exploit spatial information encoded in such documents. They extract text and perform a text-based *semantic* match to identify some metadata in the document. Our invention is superior to those as it uses a combination of text-based match and spatial

SIGNATURE INVENTOR	1/27/2000 DATE	PATENT DOCKET NO.
SIGNATURE INVENTOR	2/10/2000 DATE	
SIGNATURE INVENTOR	2/10/2000 DATE	
SIGNATURE INVENTOR	DATE	

READ AND UNDERSTOOD BY:

Daniel J Van Buer WITNESS NAME (TYPE)	Daniel J Van Buer SIGNATURE	2/21/2000 DATE
Wensheng Zhou WITNESS NAME (TYPE)	Wensheng Zhou SIGNATURE	2/21/2000 DATE

(EACH PAGE UPON WHICH INFORMATION IS ENTERED SHOULD BE SIGNED AND WITNESSED)

reasoning that better match human behavior. To the best of our knowledge, no other system is able to automatically extract the full range of metadata from scientific papers saved in PostScript format as our invention does.

CiteSeer [3] automatically generates a citation index from a set of papers. It provides a framework for literature retrieval by following citation links, evaluation of papers based on the number of citations, and identification of research trends. CiteSeer locates, downloads and parses Postscript files to extract citations from the papers in order to produce the citation index. However, CiteSeer does not extract other useful information such as title, authors, affiliations, and the like.

BIRD (BIbliometric Retrieval of Documents) [4] is a bibliometric query by example search engine. Given a set of pages of interest to the user, it retrieves a set of similar documents by following citation paths that pass through the given documents. It defines and computes a similarity measure, *relatedness*, between related and given set of documents based on the number and nature of citation linkages.

HRL LABORATORIES PROPRIETARY

SIGNATURE INVENTOR — DATE 1/27/2000

SIGNATURE INVENTOR — DATE 2/10/2000

SIGNATURE INVENTOR — DATE 2/10/2000

SIGNATURE INVENTOR — DATE

READ AND UNDERSTOOD BY:

Darrel J Van Buer
WITNESS NAME (TYPE) — SIGNATURE — DATE 2/21/2000

Wensheng Zhou
WITNESS NAME (TYPE) — SIGNATURE — DATE 2/21/2000

PATENT DOCKET NO.

(EACH PAGE UPON WHICH INFORMATION IS ENTERED SHOULD BE SIGNED AND WITNESSED)

9. DETAILED DESCRIPTION

DESCRIBE YOUR INVENTION IN DETAIL, USING NECESSARY ADDITIONAL SHEETS

A. BE SURE THAT EACH SHEET IS DATED, AND SIGNED BY EACH INVENTOR AND TWO WITNESSES. (HRL FORM 236^{C-6} CS SHOULD BE USED, IF PRACTICAL).

B. ATTACH COPIES OF DRAWINGS OR DETAILED REPORTS HELPFUL IN UNDERSTANDING HOW YOUR INVENTION WORKS

C. IF YOUR INVENTION HAS BEEN TESTED, BRIEFLY SUMMARIZE THE TEST RESULTS WHICH CONFIRM THE FUNCTIONS AND ADVANTAGES LISTED IN 8 B ABOVE.

A fundamental step in automatically ingesting and introducing electronic documents into a digital library system is to disaggregate each document into its basic constituents to allow it to be effectively indexed, searched, and disseminated. In a scientific paper, *metadata* such as author(s), affiliation(s), title, abstract, and citations play a fundamental role in consolidating the knowledge of its reader [1]. Therefore, it is important to extract such metadata in an efficient and accurate manner.

In the past, various systems have been presented to disaggregate text-based documents. They broadly fall into one of the following two categories:

- *Context-free grammar parsing.* In such an approach a somewhat rigid syntactical structure of the document is necessary. The text is composed of set of *tokens* and a set of *syntactical rules* to express legal relationships among tokens. This is the *de facto* approach for computer language interpreters and compilers. This approach requires a well-defined syntax and it is too rigid to parse free text.

- *Domain semantics based parsing.* This is an extension of the previous approach. Here a parser that embeds specific domain knowledge is used. Such a parser recognizes keywords and structural relationships for a well-defined domain of the document at hand. (A successful example of this approach for medical domains is discussed in [5].) In this class of text interpretation, the parser is highly trained to work on a specific domain and its application to another domain requires radical changes to the parser itself.

Our invention presents a *spatial knowledge based* methodology to document disaggregation. Our approach can be easily integrated with the above methods to achieve improved document metadata extraction accuracy. Our approach is based on exploiting the visual/spatial knowledge humans make use of when reading a document. In general, within a document category, a certain visual layout can be identified for all documents within that category. Think for instance of a scientific paper, you know that:

- The title is located on the *upper* portion of the *first page* and it is printed using the *largest font* on the *first page*;
- Authors are listed *immediately under* the title in some order;
- Affiliations *follow* the authors' list;
- If *only one* affiliation appear then *all* authors are associated with it;
- The *same font* is used for all authors and, similarly, for all affiliations;
- The first level headers use a larger font than the second level;
- and so forth.

In the above statements, underlined words represent metadata in our document while italic words denote relationships—spatial and of other types—among metadata and/or metadata properties. We used a rule-based language to encode the

SIGNATURE INVENTOR 1/27/2000
DATE

SIGNATURE INVENTOR 2/10/2000
DATE

SIGNATURE INVENTOR 2/10/2000
DATE

SIGNATURE INVENTOR
DATE

READ AND UNDERSTOOD BY:
Darrel J Van Buer
WITNESS NAME (TYPE) SIGNATURE 2/21/2000
DATE

Wensheng Zhou
WITNESS NAME (TYPE) Wensheng Zhou
SIGNATURE 2/21/2000
DATE

PATENT DOCKET NO.

visual knowledge in our system. As already mentioned, different "types" of document require different domain . knowledge—e.g., an article on a weekly magazine has a quite different underlying structure than a scientific paper. The knowledge base we developed so far copes with scientific papers appearing on conference proceedings and specialized journals.

Architecture

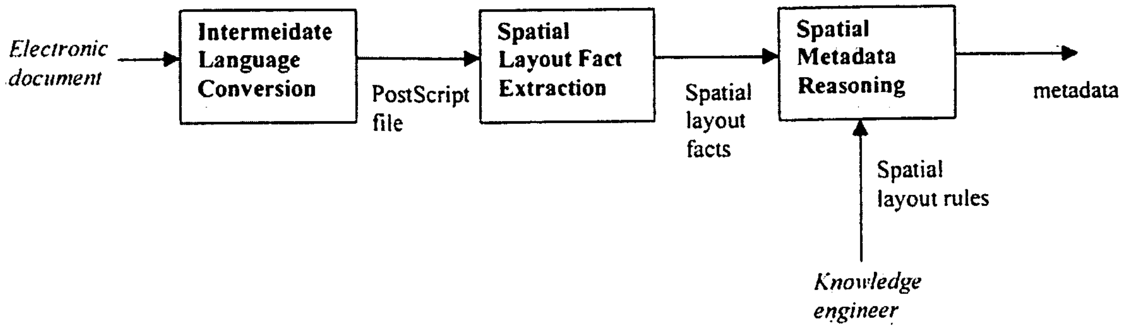


Figure 1: System Architecture

Figure 1shows the overall architecture of our system. To accommodate the diversity of file formats in which electronic documents can be encoded, we choose to use the PostScript language as a universal intermediate encoding language. An **intermediate language conversion** step is responsible for converting electronic documents into Postscript files capturing the spatial and visual aspects for document representation. This can generally be achieved by " printing the original document to a file" from the appropriate viewer of the document.

A PostScript document has to undergo a **spatial layout fact extraction** process to extract relevant spatial layout information and eliminate irrelevant information from it in preparation for further processing. Converting PostScript to other formats is not a trivial task [10] due to the nature of PostScript itself. In fact, PostScript is just a programming language—even though specifically tailored to model page layouts. As a programming language, there are many different ways to produce the same output, that is, there are many different PostScript programs that can be written to produce the same results. Thus, printing a PostScript file means *interpret-and-execute* the PostScript program contained in that file. This is a task generally accomplished by any printer Postscript driver or Postscript viewer such as " ghostview". This is a quite different task from converting a graphic format to another one, where a single *mapping* function between the two suffices.

Our system is centered on a *rule based language* called GCLIPS [6] that is used to encode spatial layout facts in documents as well as rules that reason these facts to extract metadata from them. GCLIPS is oriented to support development of *graphical* expert systems; it is built on top of CLIPS [7]. GCLIPS was recently exploited in a couple of successful knowledge based applications [8][9].

SIGNATURE INVENTOR — DATE 1/27/2000
SIGNATURE INVENTOR — DATE 2/10/2000
SIGNATURE INVENTOR — DATE 2/10/2000
SIGNATURE INVENTOR — DATE

PATENT DOCKET NO.

READ AND UNDERSTOOD BY:
Darrel J VanBuer — WITNESS NAME (TYPE) — SIGNATURE — 2/21/2000 DATE
Wensheng Zhou — WITNESS NAME (TYPE) — Wensheng Zhou SIGNATURE — 2/21/2000 DATE

We developed pstogclips by extending pstotext [11]: a PostScript-to-text translator developed at Digital as part of the *Virtual Paper* project [12]. pstotext reads a PostScript file and extracts all textual information contained in it. In a PostScript file, the typical word boundary based structure of a document is broken down into fragments (of words); no specific information is encoded in the file on how to recover the original structure from such a collection of fragments. pstotext—as well as other PostScript-to-text translators [13]—follows some heuristics to rebuild the word based structure of the original document. Furthermore, pstotext performs a set of other smart tasks such as merging hyphenated lines. However, it overlooks all spatial and font-related data contained in the input PostScript file—it simply does not need those to accomplish its task. Conversely, pstogclips retains all such additional data: its output consists of a set of *augmented* strings of text. These additional data are summarized in the following:

- Page of the document where the specified string appears;
- Absolute line counter order for each generated string;
- x-y location of the lower left corner of the string bounding box (in paper-dot coordinate systems);
- x-y location of the upper right corner of the string bounding box (in paper-dot coordinate systems);
- Font metrics (bounding-box extensions) used to represent the given string of text.

After spatial layout facts have been extracted from a PostScript file, we perform **spatial metadata reasoning** on them. The knowledge engineer (KE) provides a set of *rules* that embodies the *expertise* to extract the metadata of interest from the input document. pstogclips reads the input PostScript file and produces a set of *facts* for GCLIPS. Each fact contains a piece of information—text and spatial data—about the input PostScript document. Rules provided by KE *reason* on the extracted facts to identify (and extract) relevant metadata from the input documents.

Our knowledge base reasons on the facts extracted by pstogclips. We encoded the knowledge base by means of GCLIPS rules. We designed the rule set to extract the following information from the PostScript file: title, author(s), affiliation(s), mapping(s) author-affiliation, and table of contents. At this time, we have implemented knowledge for scientific papers appearing on conference proceedings and/or journals. The knowledge base is composed of 77 rules for a total of 788 lines of GCLIPS source code. The following table shows the GCLIPS rule usage distribution for the different extraction purposes:

Extrac. Purpose	*# of Rules Involved*
Title	9
Author(s)	12
Affiliation(s)	10
Auth.-Affil.	10
Table of Cont.	8
Print results.	19
Init and misc.	9

A fundamental component of our knowledge base is the implicit *fuzziness* involved in the visual/spatial based metadata recognition process. For instance, with reference to the list of mental activities earlier discussed, notice that:

- not always the title is printed by using the largest font on the first page,
- not all papers use numbered section headers and not always use different fonts for different section levels,
- sometimes authors are all listed on the same line next to each other while other times they are scattered across different lines,

SIGNATURE INVENTOR 1/27/2000 DATE
SIGNATURE INVENTOR 2/10/2000 DATE
SIGNATURE INVENTOR 2/10/2000 DATE
SIGNATURE INVENTOR DATE

PATENT DOCKET NO.

READ AND UNDERSTOOD BY:
Darrel J Van Buer WITNESS NAME (TYPE) SIGNATURE 2/21/2000 DATE
Wensheng Zhou WITNESS NAME (TYPE) SIGNATURE 2/21/2000 DATE

(EACH PAGE UPON WHICH INFORMATION IS ENTERED SHOULD BE SIGNED AND WITNESSED)

- when authors belong to different affiliations different methods are employed to specify their correspondence. Two of the most popular are: (1) superscripts on authors and affiliations; (2) each author is spatially closer to his/her affiliation. Still, many other different cases exist such as reporting affiliations as footnotes or listing authors vertically with respective affiliation on the right on the same line.

These exceptions represent the hardest part of the artificial visual recognition process. We coded the GCLIPS rules in our knowledge base in order to be tolerant against many of such exceptions. In some cases we develop a sort of general fuzzy strategy for certain metadata, whereas other cases have been treated as *special cases*. In the next section we further discuss it by means of real examples.

Experiment Results and Discussions

In this section we first present an example of execution of our system, we then discuss a couple of cases where a metadata extraction not based on visual/spatial properties would hardly succeed, then we turn to the accuracy of our system.

Consider the (portion of) paper of Figure 2. Once pstogclips has extracted all necessary facts from the PostScript file, we can process them from GCLIPS. The output of the GCLIPS screen for this paper is the following:

```
FILE: sigmod98
TITLE: Exploratory Mining and Pruning Optimizations of Constrained
Associations Rules
AUTHOR: Laks V.S. Lakshmanan (7)
AUTHOR: Jiawei Han (10)
AUTHOR: Raymond T. Ng (4)
AUTHOR: Alex Pang (4)
AFFILIATION 4: University of British Columbia
AFFILIATION 7: Concordia University
AFFILIATION 10: Simon Fraser University
---Table of Contents---
1  Introduction
3  Constrained Association Queries
4  Optimization Using Anti-Monotone
5  Optimization Using Succinct
6  Algorithms for Computing
6.1  Algorithm Apriori +
6.2  Algorithm Hybrid(m)
6.3  Algorithm CAP
8  Conclusions and Future Work
```

The title has been assembled from two lines into a single one. Authors have been correctly identified and linked to each respective affiliation (the index following each author's name links him/her to the affiliation). Notice that the system reported only once the affiliation "University of British Columbia" of two distinct authors. The table of contents misses some entries.

Title Extraction

At a first thought, one may think that the title of a scientific paper is contained in the first line of text (or couple of lines for longer titles) of the paper; therefore, a text based extraction from a PostScript file could be easily applied. Unfortunately, this is not the case when, for instance, authors report information on the proceedings containing the paper

SIGNATURE INVENTOR 1/27/2000 DATE

SIGNATURE INVENTOR 2/11/2000 DATE

SIGNATURE INVENTOR 2/10/2000 DATE

SIGNATURE INVENTOR DATE

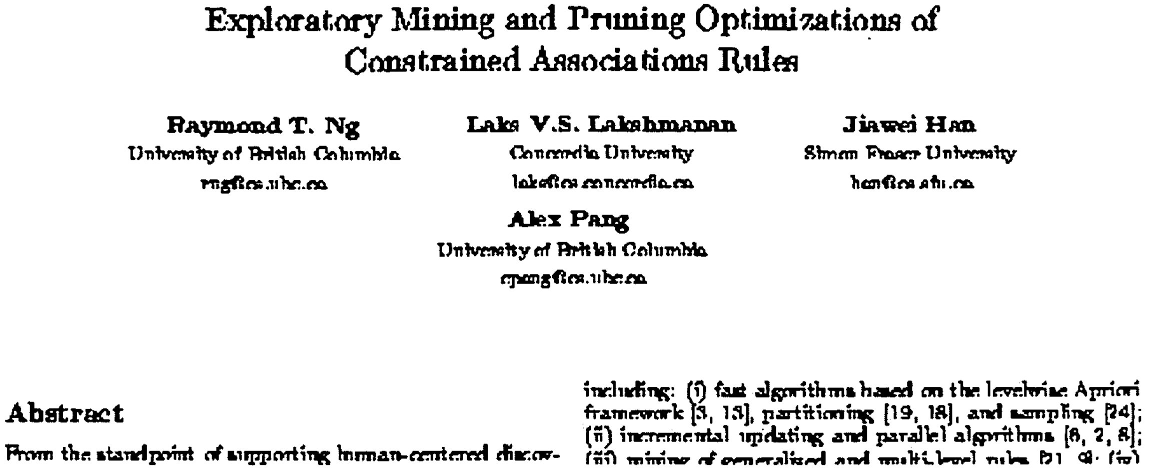
PATENT DOCKET NO.

READ AND UNDERSTOOD BY:

Darrel J Van Buer
WITNESS NAME (TYPE) SIGNATURE 2/21/2000 DATE

Wenshing Zhou
WITNESS NAME (TYPE) Wenshing Zhou SIGNATURE 2/21/2000 DATE

as shown in Figure 3. In such cases, a straight text based approach will have hard time in extracting the wanted information.



Exploratory Mining and Pruning Optimizations of
Constrained Associations Rules

Raymond T. Ng
University of British Columbia
rng@cs.ubc.ca

Laks V.S. Lakshmanan
Concordia University
laks@cs.concordia.ca

Jiawei Han
Simon Fraser University
han@cs.sfu.ca

Alex Pang
University of British Columbia
cpang@cs.ubc.ca

Abstract

From the standpoint of supporting human-centered discov-

including: (i) fast algorithms based on the levelwise Apriori framework [3, 13], partitioning [19, 18], and sampling [24]; (ii) incremental updating and parallel algorithms [6, 2, 8]; (iii) mining of generalized and multi-level rules [21, 9]; (iv)

Figure 2: Upper portion of a scientific paper

We encoded the following two hints in our knowledge base when extracting titles: (1) titles appear on the first page of the paper and (2) very often they are printed using the largest font on the first page. However, we have soon found out that not always titles are printed by using the largest font as, for instance, at times section headers use a larger (or same size) font of the title. In such a case we rely on the word "Abstract" and extract the lines printed by using the largest font among all the lines above that word. We now discuss this particular case in more details. The following GCLIPS rules are used to extract the title from the paper when the word "Abstract" was found on the first page as a stand-alone string:

```
(defrule CandidateTitleLines
  (declare (salience 9100))
  (abstract-word-found ?la)
  (doc (page 1) (font ?f $?)
       (absline ?n&:(< ?n ?la)) (text ?s))
  (metrics (page 1) (font ?f) (bbh ?hl))
  =>
  (assert (candidate-title-line ?n ?hl ?f ?s)))

(defrule GetLargestFontForCandidateTitle
  (declare (salience 9090))
  (abstract-word-found ?la)
  (candidate-title-line ?n ?hl ?f ?)
  (not (candidate-title-line ?
                             ?h2&:(> ?h2 ?hl)
                             ? ?))
  =>
  (assert (ltf ?f)))

(defrule GetTitle1
```

SIGNATURE INVENTOR 1/27/2000 DATE

SIGNATURE INVENTOR 2/10/2000 DATE

SIGNATURE INVENTOR 2/10/2000 DATE

SIGNATURE INVENTOR DATE

READ AND UNDERSTOOD BY:

Darrel J Van Buer WITNESS NAME (TYPE) Darrel J Van Buer SIGNATURE 2/21/2000 DATE

Weisheng Zhou WITNESS NAME (TYPE) Weisheng Zhou SIGNATURE 2/21/2000 DATE

PATENT DOCKET NO.

```
    (declare (salience 9000))
    (abstract-word-found ?la)
    (ltf ?f)
    (candidate-title-line ?n ?hl ?f ?s)
    (not (candidate-title-line ?n2&:(< ?n2 ?n)
                               ? ?f ?))
    =>
    (assert (paper-title ?n ?s)))

  (defrule GetTitleNextLines
    (declare (salience 9000))
    (abstract-word-found ?la)
    (ltf ?f)
    ?indx <- (paper-title ?n ?s)
    (candidate-title-line ?n2&:(= (+ 1 ?n) ?n2)
                          ? ?f ?t)
    =>
    (retract ?indx)
    (bind ?s (str-cat ?s " " ?t))
    (assert (paper-title ?n2 ?s)))
```

The first rule, `CandidateTitleLines`, asserts all lines above the one containing the word "Abstract" as candidates for the title—these will include all authors, affiliations, etc. At the same time it extracts the font size of each text line (the font size is specified in the slot bbh of the fact `metrics`). Subsequently, the rule

`GetLargestFontForCandidateTitle` extracts the largest font among all candidate title lines. The rule `GetTitle1` get the first line of the title, that is, the one that has the largest font *and* does not have any other line above it with the same font. The last rule, `GetTitleNextLines`, fires for multi-line titles, it merges successive title lines having the same font.

However, sometime this may still not be sufficient, for example, when authors names are printed using the same title font: they both appear above the abstract. Thus, we further reinforced our knowledge base by relying on the line-space (measured along the y coordinate) of title lines and authors' line.

Spatial Based Mapping Authors to Affiliation

An important metadata of a scientific paper is the affiliation of each author. Our rule base first extracts both authors and affiliations then tries to link them. There are many different cases to be considered since this is a n-to-m mapping. The simplest case is the n-to-1, in this case all n authors are affiliated to the same institute; one simple GCLIPS rules takes care of that. Another case is when the number of authors is different from the number of affiliations and there is more than one affiliation. In such a case a common practice is to use superscripts over authors and affiliations. We exploit a text-based parsing to resolve the associations in this case.

SIGNATURE INVENTOR 1/27/2000 DATE

SIGNATURE INVENTOR 2/10/2000 DATE

SIGNATURE INVENTOR 2/10/2000 DATE

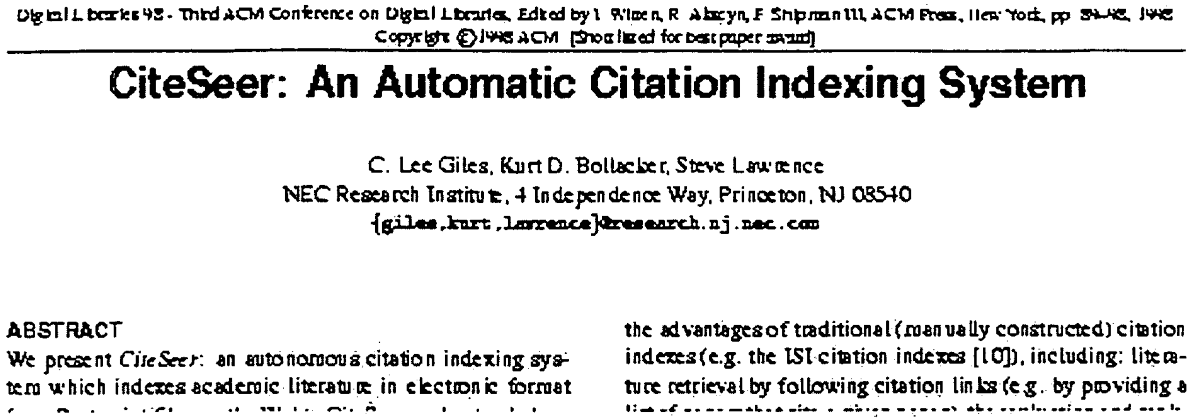
SIGNATURE INVENTOR DATE

PATENT DOCKET NO.

READ AND UNDERSTOOD BY:

Darrel J Van Buer
WITNESS NAME (TYPE) SIGNATURE 2/21/2000 DATE

Wensheng Zhou
WITNESS NAME (TYPE) SIGNATURE 2/21/2000 DATE



Digital Libraries 98 - Third ACM Conference on Digital Libraries, Edited by I. Witten, R. Akscyn, F. Shipman III, ACM Press, New York, pp. 89-98, 1998
Copyright © 1998 ACM [Shortlisted for best paper award]

CiteSeer: An Automatic Citation Indexing System

C. Lee Giles, Kurt D. Bollacker, Steve Lawrence
NEC Research Institute, 4 Independence Way, Princeton, NJ 08540
{giles,kurt,lawrence}@research.nj.nec.com

ABSTRACT
We present *CiteSeer*: an autonomous citation indexing system which indexes academic literature in electronic format

the advantages of traditional (manually constructed) citation indexes (e.g. the ISI citation indexes [10]), including: literature retrieval by following citation links (e.g. by providing a

Figure 3: Title is not the first string on the page

The case we now discuss is the n-to-n as shown in Figure 2—notice that one affiliation appears twice. In this case we perform a spatial reasoning to link each author to his/her affiliation. This is accomplished by the following GCLIPS rules:

```
(defrule XY-AffiliationLocation
  (declare (salience 5800))
  (paper-affiliations ?n ?t)
  (doc (page 1) (absline ?n) (xc ?xc) (y ?y))
  =>
  (assert (xy-affiliation ?n ?xc ?y)))

(defrule XY-AuthorLocation
  (declare (salience 5800))
  (paper-authors ?n ?t)
  (doc (page 1) (absline ?n) (xc ?xc) (y ?y))
  =>
  (assert (xy-author ?n ?xc ?y)))

(defrule SpatialLink-1
  (declare (salience 5800))
  (xy-author ?n ?xp ?yp)
  (xy-affiliation ?m ?xa ?ya)
  =>
  (assert (link-distance ?n ?m
          =(sqrt (+ (* (- ?xp ?xa) (- ?xp ?xa))
            (* (- ?yp ?ya) (- ?yp ?ya)))))))

(defrule SpatialLink-2
  (declare (salience 5800))
  (n-affiliations ?n ?)
  (n-authors ?n ?)
  (paper-authors ?na ?t)
  (not (link ?t ?))
  (link-distance ?na ?m ?d1)
  (paper-affiliations ?m ?tt)
```

HRL LABORATORIES PROPRIETARY
THIS DOCUMENT CONTAINS PROPRIETARY INFORMATION, AND EXCEPT WITH WRITTEN PERMISSION OF HRL LABORATORIES, LLC, SUCH INFORMATION SHALL NOT BE PUBLISHED, OR DISCLOSED TO OTHERS, OR USED FOR ANY PURPOSE, AND THE DOCUMENT SHALL NOT BE DUPLICATED IN WHOLE OR IN PART: THIS LEGEND SHALL BE APPLIED TO ALL DOCUMENTS CONTAINING THIS INFORMATION

SIGNATURE INVENTOR — 1/27/2000 DATE
SIGNATURE INVENTOR — 2/10/2000 DATE
SIGNATURE INVENTOR — 2/10/2000 DATE
SIGNATURE INVENTOR — DATE

PATENT DOCKET NO.

READ AND UNDERSTOOD BY:
Darrel J Van Buren — WITNESS NAME (TYPE) — SIGNATURE — 2/21/2000 DATE
Wensheng Zhou — WITNESS NAME (TYPE) — Wensheng Zhou SIGNATURE — 2/21/2000 DATE

```
(not (link-distance ?na ? ?d2&:(< ?d2 ?d1)))
=>
(assert (link ?t ?tt)))
```

The rule XY-AffiliationLocation asserts the xy location (in paper dot coordinates) of the center of the string bounding box of each affiliation (the slot xc of the fact doc contains that location). Similarly, the rule XY-AuthorLocation asserts the bounding box center xy location of each author. In turn, the rule SpatialLink-1 computes the Euclidean distance among each possible pair author-affiliation and asserts each of such possible combinations using the fact link-distance. Eventually the rule SpatialLink-2 associates each author to the (spatially) closest affiliation and asserts this by using the fact link.

Extraction of Table of Contents

When extracting table of contents we distinguish two basic cases: (1) when section headers are numbered and (2) when they are not. We use different sets of rules according to the style adopted by the paper at hand. Thus, the first thing the rule base does is to find out whether or not section headers are numbered.

Section headers numbering is a fundamental hint for a text-based extraction of table of contents; this is because the numbering is expected to follow a certain order throughout the paper and the numbers always appear at the beginning of the line. However, not infrequently, headers are not numbered, therefore an extraction based on text-parsing becomes hardly applicable. In our system we exploit the visual properties of section headers, that is, they have (1) a larger font than the text before and after and (2) a different line-space compared to the average line-space of the entire document. Furthermore, we initially look for common header names such as "Introduction," "Overview," "Motivation," or "References" to find an initial hint for the font size of the first level headers.

Performance

We tested our system over a set of 91 scientific papers randomly downloaded from the web (PostScript files). We chose our papers among conference proceedings and journal papers. (We did not include things like thesis, dissertations, or any reports with unusual layouts—we simply did not implement any expertise to read them.) We designed the rule set to extract the following information from the PostScript file: title, author(s), affiliation(s), author-to-affiliation mapping(s), and table of contents.

We ran our system on each paper and we manually checked whether or not it extracts properly the metadata of interests. The following table summarizes the overall accuracy results.

	Accuracy
Title	91%
Author(s)	86%
Affiliation(s)	74%
Auth.-Affil.	69%
Table of Cont.	75%

Partially correct results were considered as wrong, for instance, in some cases not all authors were properly identified which yielded to a negative score in our performance estimation.

SIGNATURE INVENTOR — DATE 1/27/2000

SIGNATURE INVENTOR — DATE 2/10/2000

SIGNATURE INVENTOR — DATE 2/10/2000

SIGNATURE INVENTOR — DATE

PATENT DOCKET NO.

READ AND UNDERSTOOD BY:

Darrel J Van Buer — WITNESS NAME (TYPE) — SIGNATURE — DATE 2/21/2000

Wen Sheng Zhou — WITNESS NAME (TYPE) — Wensheng Zhou SIGNATURE — DATE 2/21/2000

PD# 000211

(EACH PAGE UPON WHICH INFORMATION IS ENTERED SHOULD BE SIGNED AND WITNESSED)

"Titles" and "authors" are the most accurate findings due to their simple structural description. Mappings between authors and affiliations are difficult due to the very many different ways they are encoded. Furthermore, notice that most of the papers used in our test were never seen before, that is, we may not even have considered that specific situation during the design of our knowledge base. At this time we are already reinforcing our knowledge base to deal with them. In any knowledge based system (either natural or artificial), a knowledge refinement process takes place over time as a results of things like growing experience or trial-and-errors. We follow a similar path by reinforcing our rules over time to handle previously unseen cases.

Some errors are also due to unconventional ways of coding the PostScript file. Different drivers use different ways of coding their PostScript output. The different representations may sometime confuse the low-level PostScript feature extraction process (which may even fail its task). We need further investigation in this direction.

References

[1] A.P. Bishop. 1998. Digital Libraries and Knowledge Disaggregation: The Use of Journal Article Components, from *Digital Libraries 98.* Pittsburgh, PA, USA.

[2] *DeLIver.* http://dli.grainger.uiuc.edu/deliver.html.

[3] C. Giles and K. Bollacker and S. Lawrence. 1998. CiteSeer: An Automatic Citation Indexing System, from *Proceedings of the Third International Conference on Digital Libraries.*

[4] *Bibliometric Retrieval of Documents.* http://ai.iit.cnr.ca/.

[5] D.B. Johnson and R.K. Taira and A.F. Cardenas and D.R. Aberle. 1997. Extracting information from free text radiology reports. *International J. on Digital Libraries*, 1, pages 297--308.

[6] Giuffrida, G. and Salvemini, M. 1999 (April). Loosening The Connection Between Syntax And Semantics For Spatial Data, from *AGILE '99. Rome, Italy.*

[7] *CLIPS: A Tool for Building Expert Systems.* http://www.ghg.net/clips/CLIPS.html.

[8] G. Giuffrida and A. Vellaikal. 1999 (July). Knowledge Based Dynamic Modeling for Sensor Fusion Applications, from *NAFIPS '99.*

[9] E.C. Shek and G. Giuffrida and S. Joshi and Son K. Dao. 1999 (July). Dynamic Spatial Clustering for Intelligent Vehicle Information Sharing and Dissemination, from *SSD'99.*

[10] C.G. Nevill-Manning and T. Reed and I. H. Witten. *Extracting Text from PostScript.* Technical report. Comp. Science Dept., University of Waikato, New Zealand.

[11] *Pstotext.*http://www.research.digital.com/SRC/virtualpaper/pstotext.html.

SIGNATURE INVENTOR 1/27/2000 DATE

SIGNATURE INVENTOR 2/10/2000 DATE

SIGNATURE INVENTOR 2/10/2000 DATE

SIGNATURE INVENTOR DATE

READ AND UNDERSTOOD BY:

Darrel J Van Buer
WITNESS NAME (TYPE) SIGNATURE 2/21/2000 DATE

Wensheng Zhou
WITNESS NAME (TYPE) SIGNATURE 2/21/2000 DATE

PATENT DOCKET NO.

(EACH PAGE UPON WHICH INFORMATION IS ENTERED SHOULD BE SIGNED AND WITNESSED)

[12] *Virtual Paper.* http://research.digital.com/SRC/virtualpaper/home.html.

[13] *Prescript.* http://www.nzdl.org/technology.

HRL LABORATORIES PROPRIETARY

SIGNATURE INVENTOR 1/27/2000
DATE

SIGNATURE INVENTOR 2/10/2000
DATE

SIGNATURE INVENTOR 2/10/2000
DATE

SIGNATURE INVENTOR
DATE

READ AND UNDERSTOOD BY:

Darrel J Van Buer
WITNESS NAME (TYPE) SIGNATURE 2/21/2000
DATE

Wiensheng Zhou
WITNESS NAME (TYPE) SIGNATURE 2/21/2000
DATE

PATENT DOCKET NO.

(EACH PAGE UPON WHICH INFORMATION IS ENTERED SHOULD BE SIGNED AND WITNESSED)